http://infor.seaninstitute.org/index.php/infokum/index
JURNAL INFOKUM, Volume 7, No.2, Juni 2019 ISSN 2302-9706

TESTING K-MEANS ALGORITHM USING THE RAPID MINER APPLICATION FOR DATA GROUPING

Ade Setiawan Sembiring, Suprianto Panjaitan, Aditiya Pakpahan
[1,2,3] Program Studi Teknik Informatikan
STMIK Pelita Nusantara, Medan, Sumatera Utara Indonesia
setiawan87@gmail.com

Abstract
Data mining is a technique of extracting information that has not been known before in a collection of data in the database. Data mining has been applied in various fields that require information extraction. One of them is in grouping data. Grouping is used to divide a set of data into several useful parts so that it is easier to identify the class of data. Distribution companies can use grouping, one of which is to determine the intensity of the volume of goods ordered. This study analyzes the application of data mining with the k-means clustering algorithm to extract information from random goods ordering data. Namely by using the number of items and the total amount of the quantity of each item ordered. Then it is implemented into a website-based system to make it easier to get valid information.
Keywords: algorithmic, K-Means, rapid miner

1. Introduction

Data mining is an activity undertaken to explore information in the form of patterns or groupings of data in a data set that has a large enough amount, so that information can be taken that can be used to help make decisions. Data mining is also interpreted as extracting interesting patterns from large amounts of data. Where a pattern is said to be interesting if the pattern is not implicit, unknown beforehand, but useful. Some problems that arise in some retails are frequent packing delays due to the large number of requests for goods, but the number of personnel in the area is not qualified to do the packing for the demand for such large items. Delay in the packing of goods is also caused because the stock of goods has a large demand, but the position of the goods in the parking area is insufficient to meet the demand, so it takes time to fulfill the order of the goods. This happens because of the limited information on which items are insufficient demand.

Several methods can be used to group data. One of them is the K-Means Clustering method. K-means Clustering groups data based on certain references into several groups. where from the data each item will be grouped based on the number of orders for the item. The grouping of goods is based on the number of orders for the goods because the number of packaged goods affects the fulfillment of the order of the goods, which also relates to monitoring inventory of goods contained in the distribution center. The number of orders for goods also relates to how the utilization of personnel at each packing place.

2. Literature Riview

2.1 Data Mining

According to Tan in Eko Prasetyo's book entitled Data Mining concepts and applications using Matlab. Data mining is the process of getting useful information from a large database warehouse. Also interpreted as extracting new information, which is taken from large chunks of data that helps in decision making.

In several national journals, there are several definitions of data mining, namely:

1. Data mining is a method used to extract hidden predictive information in a database.
2. The term data mining has several views, such as knowledge discovery or pattern recognition. both of these terms have their respective accuracy.

The term knowledge discovery is because the purpose of data mining is to get knowledge that is still hidden in chunks of data. The term pattern recognition or pattern recognition remains to be used because the knowledge to be extracted is indeed in the form of patterns which also still need to be extracted from the chunks of data being faced

One technique created in data mining is how to trace existing data to build a model, and then use the model to recognize other data patterns that are not available in the stored database. The need for predictions can also utilize this technique. Besides, data mining can also be done to group data to find out the universal patterns of existing data. Anomalies and transactions also need to be detected to find out what further actions can be taken. All of these are aimed at supporting the company's operational activities so that the company's final objectives can be achieved. [1]–[4]

2.2 K-Means

K-means Clustering method is a method of group analysis that leads to the N partitioning of observation objects into K groups (clusters) where each object of observation is owned by a group with the closest mean.

K-means clustering is one method of grouping non-hierarchical data (partition) that tries to partition existing data into two or more groups. This method partitioned the existing data into groups of data, so that data with the same characteristics were entered into the same group, and data with different characteristics were put into other groups. The grouping of data aims to minimize the objective functions that are set in the grouping process, which generally tries to minimize variations within a group and maximize variation between groups. [5]–[10]

Grouping data with the K-means clustering algorithm is generally done in sequence

1. Determine k as the number of clusters you want to form
2. Initialize the initial centroid k (center point cluster) randomly.
3. Allocate each data or object to the nearest cluster. The distance between objects and the distance between objects and certain clusters are determined by the distance between the data and the center of the cluster. To calculate the distance of all data to each cluster center, one can use the ecuilidean distance theory which is formulated as the following equation.

D (i, j)

Where:

D (i, j) = distance of data i to the center of the cluster j

Xki = data to i in the attribute data to k

Xkj = cluster center j on attribute to k

The cluster center distance is recalculated with the current cluster membership. Cluster center is the average of all data or objects in a particular cluster, if desired you can also use the median value of the cluster.

4. Repeat step three until the result with the iteration is the same as the previous iteration.

There are several advantages possessed by the K-Means algorithm, namely:

1. Easier to implement and run.
2. Easy to adapt
3. Very common or often used for clustering problems.

The disadvantages of applying this method are:

1. Before the algorithm is run, k points are initialized randomly so that the resulting groupings of data can vary. If the random value for initialization is not good, the resulting groupings will be less than optimal.
2. Can get caught up in a problem called curse of dimensionality. This can happen if the training data has a very high dimension (Example if the training data consists of 2 attributes, then the dimensions are 2 dimensions. But if there are 20 attributes, there will be 20 dimensions). One way this algorithm works is to find the closest distance between k points with other points. If you are looking for distances between points in 2 dimensions, it is still easy to do. But how to find the distance between points if there are 20 dimensions. This will be difficult.
3. If there are only a few sample data points, then it is quite easy to calculate and find the closest point to k points that are initialized randomly. However, if there are many data points (for example one billion pieces of data), then the calculation and search for the nearest point will require long time. The process can be accelerated, but more complex data structures such as KD-Tree or hashing are needed.

3. Result and discussion

3.1 Data Preparation

Data used as part of the test are described in the following table:

Table 1 Sample Data Ordering Goods

DIV	MINOR	warehouse	PTAG	QTY
2	2	G009	S	100

2	2	G009		69
2	1	G009	S	93
15	5	G009		67
15	6	G009	L	32
15	10	G009		74
15	2	G009		67
24	3	G009	S	31
15	2	G009		53
10	1	G009		112
10	1	G009		97
11	2	G009		125
11	6	G009	L	67
4	4	G009		92
7	1	G009		59
7	1	G009		86
25	2	G009		46
7	1	G009		61
1	8	G009	D	131
1	2	G009		77
10	1	G009		49
25	2	G009	L	71
15	1	G009		193
3	1	G009		93

3.1 Result

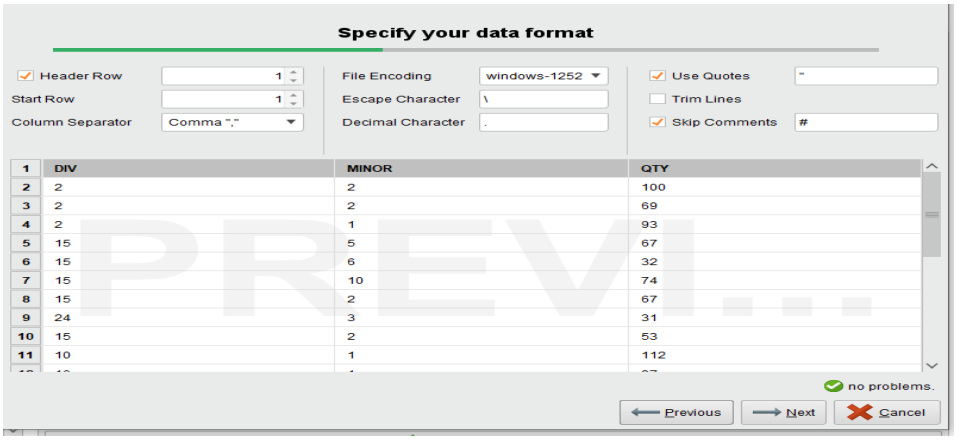
Test using the Rapid Miner application with the following steps and preparations:

1. Data Preparation Process with Microsoft Excel

	DIV	MINOR	QTY
2	2	2	100
3	2	2	63
4	2	1	93
5	15	5	67
6	15	6	32
7	15	10	74
8	15	2	67
9	24	3	31
10	15	2	53
11	10	1	112
12	10	1	97
13	11	2	125
14	11	6	67
15	4	4	92
16	7	1	59
17	7	1	86
18	25	2	46
19	7	1	61
20	1	8	131
21	1	2	77
22	10	1	49
23	25	2	71
24	15	1	193
25	3	1	93

Figure 1. Data on Excel

2. Display data on the Rapid Miner application with the following results:



	DIV	MINOR	QTY
1	2	2	100
2	2	2	69
3	2	1	93
4	15	5	67
5	15	6	32
6	15	10	74
7	15	2	67
8	24	3	31
9	15	2	53
10	10	1	112

Figure 2. Data Display in Rapid Miner

3 . K-Means Algorithm Testing Process

The process of testing with rapid miner by connecting data with k-means algorithm with the following design:

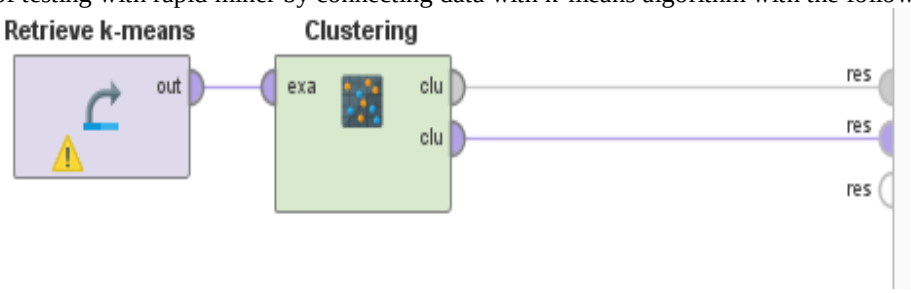


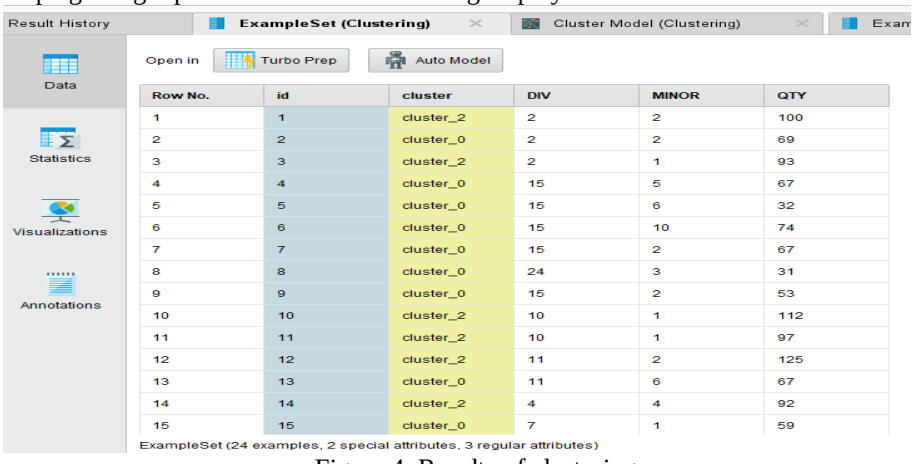
Figure 3. Clustering Design

3.2 Discussion

From the tests carried out, the results found with several displays below:

1. Clustering Results

The results of grouping using rapid miner with the following display:



Row No.	id	cluster	DIV	MINOR	QTY
1	1	cluster_2	2	2	100
2	2	cluster_0	2	2	69
3	3	cluster_2	2	1	93
4	4	cluster_0	15	5	67
5	5	cluster_0	15	6	32
6	6	cluster_0	15	10	74
7	7	cluster_0	15	2	67
8	8	cluster_0	24	3	31
9	9	cluster_0	15	2	53
10	10	cluster_2	10	1	112
11	11	cluster_2	10	1	97
12	12	cluster_2	11	2	125
13	13	cluster_0	11	6	67
14	14	cluster_2	4	4	92
15	15	cluster_0	7	1	59

Figure 4. Results of clustering

2. Statistical value

Statistical values are generated with the display below:

Name	Type	Missing	Statistics		
id	Integer	0	Min 1	Max 24	Average 12.500
cluster	Nominal	0	Least cluster_1 (1)	Most cluster_0 (14)	Values cluster_0 (14), clust
DIV	Integer	0	Min 1	Max 25	Average 10.500
MINOR	Integer	0	Min 1	Max 10	Average 2.792
QTY	Integer	0	Min 31	Max 193	Average 81.042

Figure 5. Statistical Value

3. Graphic Grouping

Graph of Grouping with the display below:

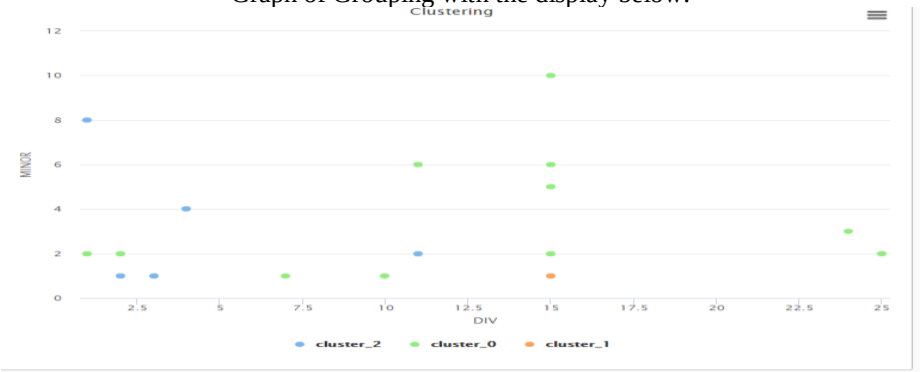


Figure 6. Grouping

4. Conclusion

The application of k-means clustering method in grouping data ordering goods can help analyze the data ordering goods that existed before. Information obtained from the application of the k-means clustering method can be used to help make decisions to overcome problems that have occurred previously

Reference

[1] S. Džeroski, "Data Mining," in *Encyclopedia of Ecology, Five-Volume Set*, 2008.
[2] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.
[3] S. Agarwal, "Data mining: Data mining concepts and techniques," in *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*, 2014.
[4] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, 2014.
[5] A. Coates and A. Y. Ng, "Learning feature representations with K-means," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 2012.
[6] D. Sculley, "Web-scale k-means clustering," in *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, 2010.
[7] G. Tzortzis and A. Likas, "The MinMax k-Means clustering algorithm," in *Pattern Recognition*, 2014.
[8] K. R. Žalik, "An efficient k′-means clustering algorithm," *Pattern Recognit. Lett.*, 2008.
[9] L. B. Neuristique and Y. Bengio, "Convergence Properties of the K-Means Algorithms," *Adv. Neural Inf. Process. Syst.*, 1995.
[10] T. M. Kodinariya and P. R. Makwana, "Review on determining number of Cluster in K-Means Clustering," *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, 2013.

